Mastering Pandas for Data Analysis

By

Yasir Hussein Shakir

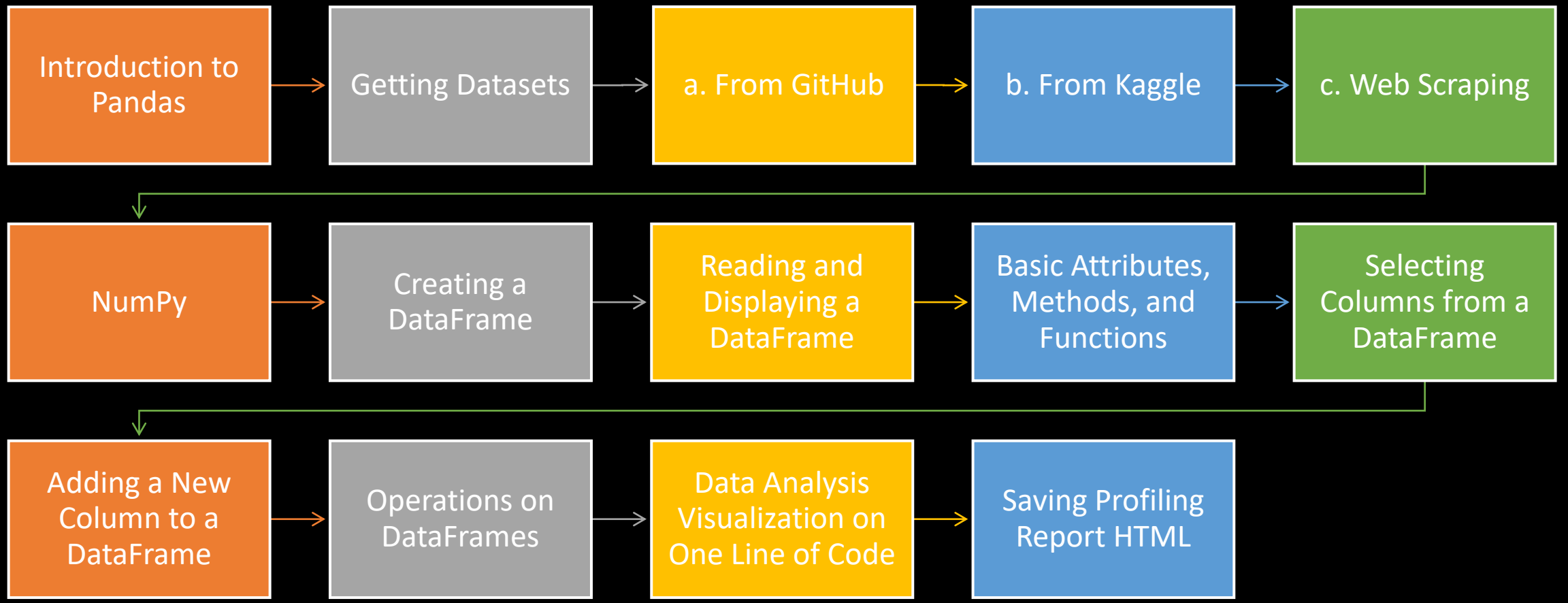
[Yasir Hussein Shakir | Master | Kaggle]

[(12) Yasir Hussein Shakir – YouTube]

[yasserhessein (Yasir Hussein Shakir) · GitHub]

[Yasir Hussein - A - yasir | LinkedIn]

Agenda


Introduction to Pandas
Getting Datasets
a. From GitHub
b. From Kaggle
c. Web Scraping
NumPy
Creating a DataFrame
Reading and Displaying a DataFrame
Basic Attributes, Methods, and Functions
Selecting Columns from a DataFrame
Adding a New Column to a DataFrame
Operations on DataFrames
Data Analysis Visualization on One Line of Code
Saving Profiling Report HTML

Introduction to Pandas

Pandas is a popular open-source Python library designed for data manipulation and analysis. It provides flexible data structures, such as Series and DataFrame, that make working with structured data seamless. Developed by Wes McKinney, Pandas is built on top of NumPy, making it an integral part of the data science ecosystem.

Why Pandas Importance

Several reasons:

- Data Manipulation: Pandas simplifies the process of cleaning, transforming, and analyzing data.
- Data Exploration: Pandas offers versatile functions for exploring and summarizing datasets, including descriptive statistics, frequency analysis, and visualizations.
- Efficient Data I/O: Whether working with CSV, Excel, SQL databases, or other formats, Pandas simplifies data input and output operations, facilitating seamless interaction with various data sources.

Getting Datasets

Name	Website
Kaggle	https://www.kaggle.com/datasets
Google Dataset Search	https://datasetsearch.research.google.com/
UCI Machine Learning Repository	https://archive.ics.uci.edu/
Papers with Code	https://paperswithcode.com/datasets
Github	https://github.com/
Data.Gov	https://catalog.data.gov/dataset
Earth Data	https://www.earthdata.nasa.gov/
NLP Index	https://index.quantumstat.com/#dataset

Web Scraping

- BeautifulSoup
- Uniform Resource Locator(URL)

a. Kaggle
b. Github
c. Web

What is the NumPy

which stands for Numerical Python, is a powerful open-source library in Python for numerical and mathematical operations.

It provides support for large, multi-dimensional arrays and matrices, along with a collection of high-level mathematical functions to operate on these arrays.

Thank you for listening

Any questions